CLAIMS

1. A method for analyzing and processing documents, comprising the steps of:

building a dictionary based on keywords from an entire text of the documents,

analyzing text of the documents for the keywords or a number of occurrences of the keywords and a context in which the keywords appear in the text; and

clustering documents into groups of clusters based on information obtained in the analyzing step, wherein each cluster of the groups of clusters includes a set of documents containing a same word or phrase.

2. The method of claim 1, wherein the clustering step clusters the documents in a catalog tree.

3. The method of claim 1, wherein the clustering step is a static clustering that does not change in response to a user query.

4. The method of claim 1, further comprising the step of splitting the groups of clusters into subclusters, the splitting step including:

finding words which are representative for each of the group of clusters;

generating a matrix containing information about occurrences of the top words in the documents from the groups of clusters; and

creating new clusters based on the generating step which corresponds to the top words and a set of phrases.

5. The method of claim 1, wherein the analyzing step includes analyzing the documents for statistical information including word occurrences, identification of relationships between words, elimination of insignificant words and extraction of word semantics.

6. The method of claim 1, wherein the clustering step is performed recursively.

7. The method of claim 1, wherein the analyzing and clustering steps are performed off line.

8. The method of claim 1, further comprising the step of generating specific tags for the documents including at least one of document title, document language and summary and the keywords.

9. The method of claim 1, further comprising the step of assigning weights to the words and computing the appropriate weights of sentences within the documents.

10. The method of claim 1, further comprising the step of summary generation of the documents, the summary generation being based on the assigned weights to the words and the appropriate weights of the sentences.

11. The method of claim 1, wherein the analyzing step is performed on only selected documents which are marked.

12. The method of claim 11, wherein the documents are HTML documents.

13. The method of claim 12, wherein the analyzing step includes applying linguistic analysis to the documents, the linguistic analysis being performed on one of titles, headlines and body of the text, and content including at least one of phrases and the words.

14. The method of claim 13, wherein the dictionary generates words that describe the contents of the documents, creates indexes for the documents, associates the documents with other documents to create concept hierarchy, clusters the documents using a tree-structure of the concept hierarchy and generates a best-suited phrase for cluster description.

15. The method of claim 14, wherein the dictionary includes all words appearing in the analyzed documents, and the documents are indexed with the words from the dictionary.

16. The method of claim 15, wherein importance is assigned to each word in the document, the importance being a function of word appearances in the document, position in the document and occurrences in links pointing to the document.

17. The method of claim 1, further comprising detecting a language of the documents based on frequencies of letter occurrences and co-occurrences in the words.

18. The method of claim 1, wherein the clustering step is based on one of (i) a best-suited phrase or word from the documents and (ii) generation word conjunction templates for grouping the documents.

19. The method of claim 1, wherein the analyzing step includes extracting document meta information.

20. The method of claim 1, further comprising the steps of

generating a cluster heirarchy for the groups of clusters;

genering cluster descriptions, the clustering descriptions including words or phrases that generate a cluster of the groups of clusters and the number of the documents in the cluster; and

assigning the documents to elementary clusters and indirect clusters.

21. The method of claim 20, wherein a cluster of the groups of clusters is split into subclusters using statistics to identify best parent cluster and most discriminating significant word in the cluster.

22. The method of claim 1, further comprising the step of processing the documents, the processing including:

creating reverted index of occurrences of words and phrases in the documents;

building a directed acyclic graph; and

extracting a limited number of representative sentences or words or phrases for the document.

23. The method of claim 21, wherein the processing step is independent of the clustering step and is performed in incremental steps.

24. The method of claim 23, wherein the clustering step includes the steps of:

creating reverted index of occurrences of words and phrases in the documents;

building a directed acyclic graph; and

counting the documents in each group of clusters.

25. The method of claim 24, wherein the clustering step further includes:

generating document summaries and statistical data for the groups of clusters;

updating global data by using the document summaries;

generating cluster descriptions of the groups of clusters by finding representative documents in the each cluster of the groups of clusters;

finding elementary clusters associated with the groups of clusters which contain more than a predetermined size of the documents; and

storing the elementary clusters in storage.

26. The method of claim 1, wherein the analyzing step includes transforming unstructured textual data associated with the documents into structured data in form of tables.

27. The method of claim 1, wherein the analyzing step includes the steps of:

computing a basic weight of a sentence as a sum of weights of the words in the sentence;

normalizing the weight with respect to a length of the sentence;

selecting sentences with highest weights;

ordering the sentences with the highest weights in an order which they occur in the input text;

providing a priority to the words by evaluating a measure of particular occurrence of the words in the documents; and

extracting the keywords from the documents which are representative for a given document, the keywords being extracted as follows:

for each word **s** occurring in the document **D**

compute an importance index for s using the formula:

$$\textbf{Importance(s,D)} =$$

$$= [\textbf{Priority(s,D)}/\text{size}(\textbf{D})] \log[N/DF(\textbf{s})]$$

where N is a number of all the documents and DF(s) is the number of all the documents which contain the word s.

28. The method of claim 1, wherein the documents are divided into different topic domains and restricted to document size.

29. The method of claim 28, wherein a critical size of the documents is determined prior to the analyzing step such that when the critical size exceeds a predetermined size, the analyzing step only analyzes a first part and a last part of the documents.

30. The method of claim 1, wherein the analyzing step includes splitting the documents into separate lexemes including words and hypertext markup language (HTML) tags.

31. The method of claim 30, wherein the analyzing step further comprises the steps of:

determining whether there is a next lexeme in the documents;

computing the priorities of all of the words in the documents if the next lexeme is found;

determining which type of information is the lexeme; and

if the documents contain a word lexeme then:

obtain an identification of the word from the dictionary;

update statistics of the word occurrence; and

return an ID of the word.

32. A system for analyzing and processing documents, comprising the steps of:

a module for building a dictionary based on the keywords from an entire text of the documents,

a module for analyzing text of the documents for the keywords or a number of occurrences of the keywords and a context in which the keywords appear in the text; and

a module for clustering documents into groups of clusters based on information obtained in the analyzing step, wherein each cluster of the group of clusters is a set of documents containing a same word or phrase.

33. A machine readable medium containing code for analyzing and processing documents, comprising the steps of:

building a dictionary based on the keywords from an entire text of the documents,

analyzing text of the documents for the keywords or a number of occurrences of the keywords and a context in which the keywords appear in the text; and

clustering documents into groups of clusters based on information obtained in the analyzing step, wherein each cluster of the group of clusters is a set of documents containing a same word or phrase.